Biological Journal of the Linnean Society, 2017, **121**, 133–149. With 5 figures.

Statistical hybrid detection and the inference of ancestral distribution areas in *Tolpis* (Asteraceae)

MICHAEL GRUENSTAEUDL[1*], BRYAN C. CARSTENS[2], ARNOLDO SANTOS-GUERRA[3] and ROBERT K. JANSEN[4,5]

[1]*Institut für Biologie-Botanik, Dahlem Centre of Plant Sciences, Freie Universität Berlin, Altensteinstrasse 6, 14195 Berlin, Germany*
[2]*Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, OH 43210, USA*
[3]*Calle Guaidil No. 16, CP 38280, Tegueste, Tenerife, Islas Canarias, Spain*
[4]*Department of Integrative Biology, University of Texas at Austin, 1 University Station C0930, Austin, TX 78712, USA*
[5]*Biotechnology Research Group, Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Received 23 August 2016; revised 28 September 2016; accepted for publication 7 October 2016

Many historical biogeographic studies do not account for the effect of hybrid taxa on phylogenetic tree inference, despite recent advances in the statistical identification of such taxa. This investigation aims to illustrate the impact that hybrid taxa can have on ancestral area reconstructions of the plant genus *Tolpis*, which displays an evolutionary history possibly indicative of a back-colonization of the continental Mediterranean. We evaluate and apply two statistical hybrid detection methods, JML and STEM-hy, which assist in identifying reticulate patterns of allele coalescence. We also evaluate and apply a software tool, P2C2M, to test the fit of genetic loci to the multispecies coalescent model (MSCM). The application of these tools to a previously published data set of three nuclear DNA markers indicates the presence of several potential hybrid taxa in *Tolpis*. One nuclear marker is found to display a reduced level of reticulate history, a good fit to the MSCM and minimal signal of gene flow across archipelagoes. Ancestral distribution areas are reconstructed on gene and species trees of *Tolpis* before and after the exclusion of putative hybrid taxa using stochastic character mapping, parameterized likelihood reconstructions, and reconstructions under continuous-time Markov chain models. The results of these reconstructions indicate that taxa of hybrid origin may have a considerable impact on ancestral area reconstruction and that it is important to account for such taxa prior to biogeographic analysis. We conclude that *Tolpis* has likely had a time-consistent distribution in island habitats and originated on the Canary Islands.

ADDITIONAL KEYWORDS: ancestral area reconstruction – back-colonization – historical biogeography – hybridization – incomplete lineage sorting – island plants – Macaronesia – species trees – *Tolpis*.

INTRODUCTION

Homoploid hybrid and allopolyploid species can have significant effects on phylogenetic tree inference (Alvarez & Wendel, 2003). In gene tree estimation, DNA sequences of hybrid taxa may not only be recovered in close relation to their respective parents, but also at or near the base of the clade containing both parents, in a position basal to all ingroup taxa, or as sister to an unrelated taxon (McDade, 1992; Soltis *et al.*, 2008). Tree inference under cladistic criteria may be particularly impacted by hybrids formed by distantly related species that are included alongside their parents; their presence in a data set may distort the inference of relationships even for unrelated taxa (McDade, 1990, 1992). Furthermore, the inferred phylogeny may be less resolved than trees estimated without hybrid species (Nieto-Feliner, Fuertes-Aguilar & Rosselló, 2001; Moody & Rieseberg, 2012). Hybrid and allopolyploid species also affect analyses performed subsequently to

*Corresponding author. E-mail: m.gruenstaeudl@fu-berlin.de

© 2017 The Linnean Society of London,

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

tree inference, such as the reconstruction of ancestral character states (see Supporting Information, Fig. S1). If the tree topology in an ancestral character reconstruction displays uncertainty (e.g. due to the presence of hybrid species), the character reconstruction will also be uncertain (Pagel, Meade & Barker, 2004). This dependency necessitates that uncertainty in the phylogeny reconstruction should be accounted for prior to the character reconstruction (Ronquist, 2004).

Two strategies have often been employed to reduce the impact of hybridization among taxa on phylogenetic tree inference and subsequent analyses. Investigators have excluded suspected or known hybrids or allopolyploids from their data sets or removed DNA sequences that displayed reticulate relationships during tree inference (i.e. 'diploid-only data sets'; Clarkson *et al.*, 2004; Timme, Simpson & Linder, 2007). Alternatively, investigators have utilized direct amplicon DNA sequences instead of isolating allele sequences via cloning despite a priori knowledge of hybrid taxa among the study organisms (e.g. Tippery & Les, 2011; Toepel *et al.*, 2011). However, tree inference based on direct amplicon sequences is unlikely to recover true species relationships because they represent only a subset of the true homolog diversity (Nieto-Feliner & Rosselló, 2007). Both of these strategies are only applicable if the identity of hybrid taxa included in a data set is known *a priori*.

Recent advances in phylogeny inference and statistical hybrid detection have facilitated an integrated approach towards phylogenetic uncertainty caused by hybrid taxa. For example, the inference of species trees under the multispecies coalescent model (MSCM) can account for reticulations in the underlying gene trees (e.g. Yu *et al.*, 2014) or enable the removal of data sets that do not fit the coalescent model prior to tree inference (e.g. Gruenstaeudl *et al.*, 2016). Also, several methods for the statistical identification of hybridization events have been developed (e.g. Kubatko, 2009; Joly, 2012; Yu, Degnan & Nakhleh, 2012), which can help identify and account for the phylogenetic uncertainty introduced by hybrid or allopolyploid species. However, most of these methods have yet to be evaluated and adopted for historical biogeography, where recent investigations have not explicitly addressed hybrid and allopolyploid taxa (e.g. Zhang *et al.*, 2014; Fougère-Danezan *et al.*, 2015; Xiang *et al.*, 2015). Here, we assess the impact that taxa of hybrid origin can have on ancestral area reconstruction (AAR) and evaluate techniques that may detect their presence.

The plant genus *Tolpis* Adans. (Asteraceae) is ideally suited to illustrate the impact of hybrid taxa on AAR. *Tolpis* inhabits four of the five archipelagoes that constitute Macaronesia (Jarvis, 1980) and has presumably colonized each archipelago only once (Gruenstaeudl, Santos-Guerra & Jansen, 2013). Two species are distributed in the continental Mediterranean regions of Europe and North Africa (hereafter 'mainland'), and the genus comprises at least three species suspected of hybrid origin (Fig. 1): the diploid species *Tolpis crassiuscula* Svent. and *Tolpis succulenta* (Dryand. in Aiton) Lowe and the tetraploid species *Tolpis glabrescens* Kaemmer (Gruenstaeudl *et al.*, 2013). *Tolpis* was also suggested to have colonized the mainland from an island habitat (Moore *et al.*, 2002), a pattern that would place *Tolpis* into a distinct group of Macaronesian plant lineages that have experienced a back-colonization of the continent (Mort *et al.*, 2002; Allan *et al.*, 2004; Carine *et al.*, 2004; Caujapé-Castells, 2011).

We attempt to assess the impact that taxa of hybrid origin have on the biogeographic reconstructions of *Tolpis* by applying three types of analyses to a previously published multigene data set of the genus. First, we infer species trees of *Tolpis* following the identification and exclusion of nuclear loci that do not fit the MSCM. Second, we evaluate and apply statistical methods designed to identify taxa with a reticulate evolutionary history. Third, we reconstruct ancestral distribution areas on gene and species trees with different stochastic and model-based methods. Distribution areas are reconstructed before and after the exclusion of taxa statistically identified as potential hybrids. On the basis of these analyses, we discuss the hypothesis of a back-dispersal to the mainland by *Tolpis*.

MATERIALS AND METHODS

GENERALIZATIONS ON HYBRID SPECIATION

Different forms of plant hybridization exist, which result in different types of hybrid species (Soltis & Soltis, 2009; Gompert & Buerkle, 2016). All contain a horizontal exchange of genetic material, which generates a characteristic genetic signal that can be detected and reconstructed as a reticulate pattern of allele coalescence (Linder & Rieseberg, 2004). Most current statistical methods for the detection of such reticulations do not differentiate between the precise forms of hybridization or reliably differentiate between palaeo- and more recent hybridizations. We, therefore, collectively refer to species with reticulate patterns of allele coalescence as 'hybrid taxa' or 'taxa of hybrid origin', given the understanding that their age and genetic origin may be more diverse than can be discussed here (Gompert & Buerkle, 2016).

DNA SEQUENCE DATA

A previously published nucleotide sequence data set was employed for this investigation (Gruenstaeudl *et al.*, 2013). It comprises DNA sequence alignments

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

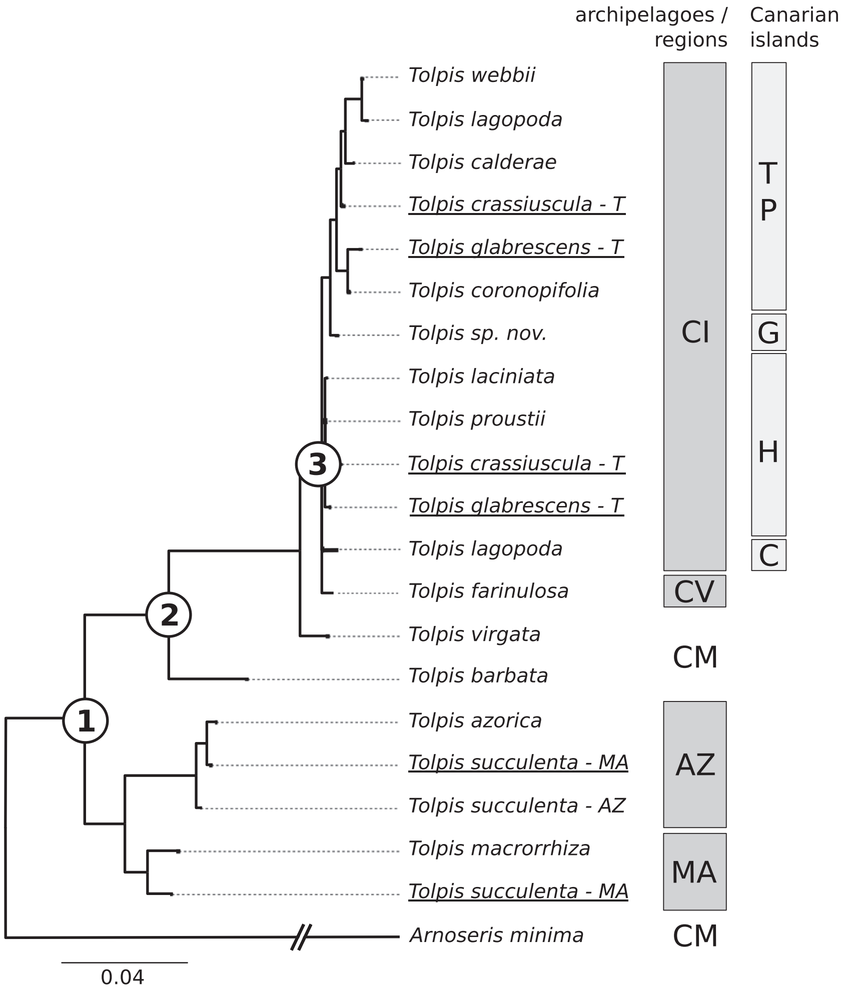

Figure 1. The maximum likelihood tree of *Tolpis* as inferred from ETS allele sequences. Circled nodes represent the MRCAs of the three focal clades. For the purpose of a clearer visualization and only in this figure, allele sequences that stem from the same species and form exclusive clades were collapsed into single terminals. Alleles of taxa suspected to be hybrids by previous investigations are underlined. Geographic distributions of extant taxa are indicated by the bars on the right. Grey bars indicate an exclusive distribution in island habitats. Archipelago abbreviations used: AZ, Azores; CI, Canary Islands; CM, Continental Mediterranean; CV, Cape Verde; MA, Madeira. Island abbreviations used: C, Gran Canaria; G, La Gomera; H, El Hierro; P, La Palma; T, Tenerife.

of two low-copy nuclear markers (A19 and B12, abbreviated 'LCNM') and the nuclear ribosomal external transcribed spacer (ETS) region. The two LCNM are collectively referred to as locus set 1 (LS1), the set of all three nuclear DNA markers as locus set 2 (LS2). Each alignment contains sequences of every recognized *Tolpis* species (Jarvis, 1980), including a novel species from the island of La Palma (Crawford, Mort & Archibald, 2013) and sequences from a potentially novel species of La Gomera (Gruenstaeudl *et al.*, 2013). The alignments include sequences from *Arnoseris minima* (L.) Schweigg. & Koerte, the putative sister taxon to *Tolpis* (Gruenstaeudl *et al.*, 2013). Each *Tolpis* species is represented by ten allele sequences per locus, which had been isolated from PCR products via TOPO TA cloning and then Sanger sequenced (Gruenstaeudl *et al.*, 2013). Alignment length, number of variable and parsimony informative sites, maximum divergence of allele sequences of selected taxa (including those of potential hybrids), and resampling support for clades of interest are provided in Table 1.

Gene and species tree estimation

To maintain consistent statistical frameworks across different hybrid detection methods, parallel sets of software programs were employed for gene and species tree estimation. In the Bayesian framework, gene tree estimation was conducted with BEAST v.1.8 (Drummond *et al.*, 2012) and species tree estimation

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

Table 1. Alignment length, number of variable sites, number of parsimony informative sites, best-fitting model of nucleotide substitution, largest *p*-distance within each taxon set, and phylogenetic resampling support for each clade of the loci under study

Locus or locus set	A19	B12	ETS	LS 1	LS 2
Alignment length	441 bp [435 bp]	420 bp [419 bp]	1338 bp [1327 bp]	n.a.	n.a.
Number of variable sites	133 (30.2%) [109 (25.1%)]	84 (20.0%) [73 (17.4%)]	502 (37.5%) [448 (33.8%)]	n.a.	n.a.
Number of PI sites	86 (19.5%) [84 (19.3%)]	61 (14.5%) [61 (14.6%)]	306 (22.7%) [296 (22.3%)]	n.a.	n.a.
Best-fitting model	HKY + G	TrN + I	GTR + G	n.a.	n.a.
Max. *p*-distance within *Tolpis*	0.184	0.112	0.192	n.a.	n.a.
Max. *p*-distance among Canarian endemics of *Tolpis*	0.176	0.087	0.121	n.a.	n.a.
Max. *p*-distance among clones of *T. succulenta* from Madeira	0.007	0.073	0.046	n.a.	n.a.
Max. *p*-distance among clones of *T. glabrescens*	0.119	0.020	0.020	n.a.	n.a.
Max. *p*-distance among clones of *T. crassiuscula*	0	0	0.018	n.a.	n.a.
Resampling support – Clade 1	BS 100; PP 1.00	BS 100; PP 1.00	BS 100; PP 1.00	BS 100; PP 1.00	BS 100; PP 1.00
Resampling support – Clade 2	BS 98; PP 1.00	BS 89; PP 1.00	BS 77; PP 0.80	BS 100; PP 0.58	BS < 50; PP < 0.50
Resampling support – Clade 3	BS < 50; PP < 0.50	BS 99; PP 0.94	BS 74; PP 0.97	BS 98; PP 1.00	BS < 50; PP < 0.50

P-distances were corrected using the best-fitting nucleotide substitution model and averaged among clone sequences. DNA sequence statistics and *p*-distances are not applicable to the marker sets, which were not concatenated. Numbers in parentheses indicate percentages of total alignment length. By default, metrics were calculated on data sets that included the outgroup; metrics calculated on data sets without the outgroup are given in square brackets. bp, base pairs; BS, bootstrap (under maximum likelihood); max., maximum; n.a., not applicable; PI, parsimony-informative; PP, posterior probability.

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

with *BEAST v.1.8 (Heled & Drummond, 2010). Markov chains were run for 50 M generations, with every 10 000th generation sampled, the first 20% of sampled generations discarded as burn-in, and every fourth generation retained during subsampling. To ensure compatibility with tests of putative hybrid taxa and model fit, a piecewise-constant model of population size change was employed. Independent sampling and convergence of Markov chain parameters was evaluated with Tracer v.1.6 (Rambaut *et al.*, 2014). Maximum clade credibility trees were generated with TreeAnnotator v.1.8 (Drummond *et al.*, 2012). In the likelihood framework, gene tree estimation was conducted with Garli v.2.0.1 (Zwickl, 2006) and species tree estimation with STEM v.2.1.0 (Kubatko, Carstens & Knowles, 2009). A set of 100 independent search replicates was employed, with every 100th generation per run sampled. Maximum clade credibility trees, the best gene, and species trees inferred under maximum likelihood, DNA sequence matrices, and specifications of the distribution areas were submitted to Treebase (http://treebase.org; submission 17247).

MSCM fit and among-archipelago gene flow

Fit to the MSCM was evaluated via posterior predictive simulation using P2C2M v.0.6 (Gruenstaeudl *et al.*, 2016). Analyses were conducted on each posterior distribution of gene and species trees using 50 simulation replicates per analysis. The coalescent likelihood function 'lcwt' was used as summary statistic. Since the relative ability of P2C2M to identify poor fit to the MSCM has yet to be evaluated for cases of hybridization, its statistical power was evaluated using simulations. Here, 20 data sets, each comprising three sequence alignments, were simulated under the same number of species and allele sequences as observed in the empirical data. Effective sample sizes of the loci were estimated with migrate-n v.3.6.6 (Beerli, 2006), average tree depths of the posterior gene tree distributions with Mesquite v.2.75 (Maddison & Maddison, 2011). Species trees were simulated under a Yule model, gene trees with hybrid-Lambda v.0.2 (Zhu *et al.*, 2015), with allele sequences of hybrid taxa shared equally between both parents. Hybridization was only specified in the third gene of each simulated data set. Nucleotide sequence data were simulated with Seq-Gen v.1.3.2 (Rambaut & Grassly, 1997) under the same substitution models and mutation rates as observed in the empirical data.

In addition to the evaluation of fit to the MSCM, we also assessed the level of incomplete lineage sorting (ILS) across and gene flow between different Macaronesian archipelagoes using Bayesian and likelihood-based hypothesis testing (Goldman, Anderson & Rodrigo, 2000). Details of these analyses are presented alongside Table S1 (see Supporting Information).

Detection of potential hybrid taxa

Two statistical methods were employed to detect potential hybrid taxa in *Tolpis*. First, the presence of reticulate patterns of coalescence was evaluated with Bayesian posterior predictive checking using minimum pairwise sequence distances as test statistic (Joly, McLenachan & Lockhart, 2009). The software implementation of this method (JML v.1.0.1; Joly, 2012) identifies species pairs that are more similar than expected under a model of pure ILS and has been applied to several plant lineages (e.g. Jones *et al.*, 2014; Joly, Heenan & Lockhart, 2014). Specifically, JML determines if ILS is a sufficient measure to explain violations of the MSCM based on a distribution of expected pairwise distances across the posterior distribution of species trees. We apply JML on each locus-specific posterior distribution of gene trees, using diploid heredity scalars and a significance level of $P = 0.05$. Second, the presence of reticulate patterns of coalescence was evaluated through model selection among hybrid species trees (Kubatko, 2009). This method can discriminate between ILS and hybridization and has been integrated into the species tree estimation software STEM (as 'STEM-hy'; Kubatko, 2009). STEM-hy maximizes the likelihood of a species tree given the probability density of gene trees that display topologies and branch lengths consistent with a shared ancestry between a hybrid and a parental taxon. Analyses in STEM-hy were conducted with three independent runs, a cooling rate of 0.005 and locus-specific scaling of branch lengths. The phylogenetic uncertainty of the species tree estimation was accounted for by conducting analyses over the 100 best species trees (Huelsenbeck, Rannala & Masly, 2000).

Prior to applying statistical hybrid detection methods to *Tolpis*, we evaluated their statistical power using several simulated and one empirical data set. The simulated data sets were parameterized to have the same number of species and allele sequences as the empirical data, but differed in the precise coalescence patterns of the alleles, with selected taxa set up as hybrids between two of the other taxa (see Supporting Information, Fig. S2). Simulation settings were identical to those used in the evaluation of the statistical power of P2C2M. The empirical data set constitutes allele sequences of diploid North American roses (Joly & Bruneau, 2006) and had been used to assess the principal applicability of JML (Joly, 2012); it allows a conservative test of hybrid detection, as it displays no or only minimal indications of allele recombination (Joly & Bruneau, 2006; Joly, 2012). It is also similar to the empirical data sets of *Tolpis* by the absence of

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

sequences of polyploid taxa. Gene and species tree distributions of the empirical data set were inferred under the same conditions as described in Joly (2012). To reduce calculation time, the species distribution of the empirical data set was subsampled to 100 post-burnin species trees.

Ancestral area reconstruction

Three types of AAR were conducted to infer the biogeographic history of *Tolpis*: (1) stochastic character mapping, (2) likelihood-based reconstruction with different dispersal rates to and from archipelagoes, and (3) reconstruction under continuous-time Markov chain (CTMC) island models with an instantaneous rate matrix reflecting geographic distance between distribution areas. The reconstructions were conducted over entire tree distributions to accommodate the uncertainty at different levels of the reconstruction process (Ronquist, 2004). For brevity, only the reconstructions of three nested focal clades were recorded (Fig. 1): the clade comprising all extant species of *Tolpis* (clade 1), the clade comprising all Canarian and continental species of the genus (clade 2), and the clade comprising all *Tolpis* species endemic to the Canary Islands (clade 3). Reconstructing the ancestral distribution areas of the most recent common ancestor (MRCA) of each of these clades is necessary to assess the general biogeographic history of *Tolpis* and discuss the hypothesis of a back-colonization. While included in all analyses, we do not explicitly discuss the Cape Verde species *Tolpis farinulosa* (Webb) J.A. Schmidt; as a member of the Canarian clade of *Tolpis* (Gruenstaeudl *et al.*, 2013), any conclusions on the Canarian species of *Tolpis* also apply to this species.

The primary method of AAR employed was stochastic character mapping (Huelsenbeck, Nielsen & Bollback, 2003), which was performed in Mesquite using the wrapper script set WARACS (Gruenstaeudl, 2016). Terminal taxa of the input phylogenies were coded by geographic location, with each Macaronesian archipelago as well as the mainland coded as separate, discrete character states. Stochastic character mapping was performed before and after the exclusion of taxa that were statistically identified as possible hybrids by STEM-hy.

Reconstructions under maximum likelihood (Pagel, 1999) utilized different rates of dispersal to and from Macaronesian archipelagoes. Five dispersal models were evaluated: a symmetrical model in which mainland-to-archipelago dispersal occurred with the same probability as dispersal in the reverse direction (M1), two asymmetrical models with mainland-to-archipelago dispersal set to be 50 and 100% more likely than a reverse dispersal (M2 and M3, respectively), and two asymmetrical models with dispersal set to be 50 and 100% less likely than a reverse dispersal (M4 and M5, respectively). Model M1 constitutes a Markov k-state 1 parameter model (Lewis, 2001) and was used to estimate dispersal rates directly. The asymmetrical models (M2–M5) constitute Markov k-state 2 parameter models. Reconstructions were performed on maximum clade credibility trees of the posterior tree distributions and compared on the basis of likelihood scores (Posada & Crandall, 2001).

For reconstructions under CTMC island models, dispersal events were modelled by an instantaneous rate matrix that defined relative distances between the distribution areas of *Tolpis* as transition probabilities (Ronquist & Sanmartin, 2011). Optimal reconstructions were inferred via MCMC sampling in MrBayes v.3.2.3 (Ronquist *et al.*, 2012) in a combined analysis of sequence data and geographic characters. Topology-constrained Markov chains were run for 20 M generations, with 50% discarded as burnin. The geographic data partition was analysed under a general time-reversible model, the DNA sequence partitions under the best-fitting models of nucleotide substitution. Dispersal probabilities were modelled under a negative exponential distribution. Dispersal rate and nucleotide sequence evolution were estimated independently with unlinked rate priors across partitions.

RESULTS

MSCM fit and among-archipelago gene flow

Simulation testing using P2C2M indicates that it can accurately detect cases of poor model fit caused by hybridization with a significance level of $P = 0.10$. However, this low rate of type II error came at the price of a high rate of type I error (Fig. 2). The selection of more stringent significance levels decreased the rate of type I and increased the rate of type II error. Under $P = 0.01$, only one out of 40 genes simulated under the MSCM exhibited poor model fit, while three genes simulated under hybridization fit the model. The application of P2C2M on the empirical data of *Tolpis* identified poor model fit for the ETS at $P = 0.01$ (Table 2). In the estimation of LS2 species trees, it was the only locus to not fit the MSCM. The LCNM, by contrast, were identified to fit the coalescent model both under LS1 and LS2 species trees.

Support for the presence of ILS across or gene flow between different archipelagoes in *Tolpis* was identified under both Bayesian and likelihood-based hypothesis testing (see Supporting Information, Table S1). The results of the AU tests showed that trees of each posterior gene tree distribution became significantly worse when they were topologically constrained to exclude among-archipelago ILS or gene flow. The same result was recovered by log Bayes factors. Species trees based

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

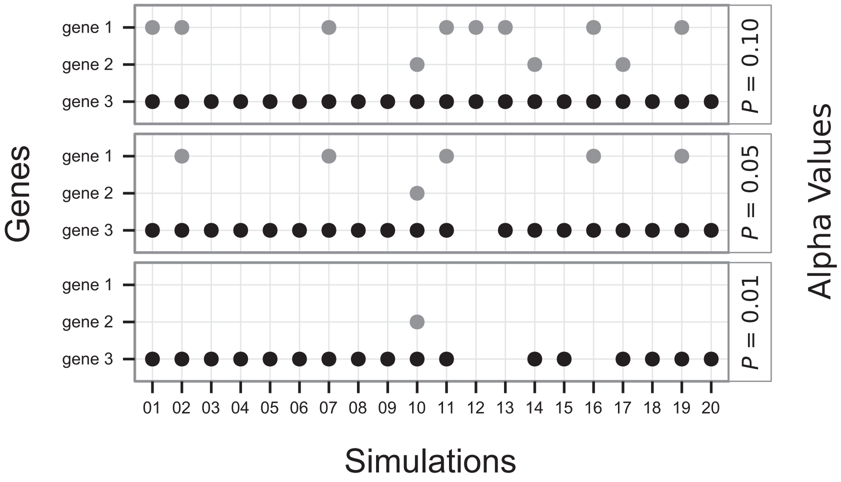

Figure 2. Results of simulations to evaluate the statistical power of P2C2M to identify poor fit to the MSCM as caused by hybridization. Each box displays locus-specific results across all simulated data sets. Different boxes display results under different significance levels (i.e., *P* = 0.10, *P* = 0.05, and *P* = 0.01). Instances of correct identification of data sets with poor fit to the MSCM are represented by black dots, instances of false positive results by gray dots.

Table 2. Results of analyses with P2C2M on empirical data of *Tolpis*

Data set	Locus	Test statistic (±1 SD)	Significance
Species tree – LS1	A19	–72.06 (±77.36)	NS
	B12	–63.61 (±69.97)	NS
Species tree – LS2	A19	6.63 (±9.11)	NS
	B12	0.81 (±11.15)	NS
	ETS	–9.75 (±4.29)	significant

Significance is inferred under *P* = 0.01. NS, not significant; SD, standard deviation.

on all three loci also displayed worse likelihoods and log Bayes factors when constrained. LS1 species trees, however, displayed a level of ILS or gene flow that was not distinguishable from homoplasy. Here, the AU test did not recover a significant difference between constrained and unconstrained trees. Moreover, the log Bayes factor for this comparison was found positive, indicating support for very low levels of ILS across or gene flow between archipelagoes. In summary, the posterior distribution of LS1 species trees was least affected by a topological constraint, implying the absence of among-archipelago ILS or gene flow. We thus considered AAR conducted on the LS1 posterior tree distribution to be more reliable than AAR on the other tree distributions.

Evaluation of statistical hybrid detection

Both hybrid detection methods indicated the presence of taxa with reticulate evolutionary histories in the simulated data sets. However, only STEM-hy exhibited an acceptable statistical specificity and sensitivity. STEM-hy, which conducts tests for hybridization under all loci collectively, correctly identified the involvement of the hybrid and one parental taxon in more than three quarters of the simulated data sets (see Supporting Information, Fig. S3). Incorrect inferences of hybridization existed, but occurred in less than one quarter of the simulations. JML, which conducts tests for hybridization in each locus independently, displayed lower specificity and sensitivity. It displayed high rates of type I error by indicating hybridization where none was modelled (17/20 under *P* = 0.05; 10/20 under *P* = 0.01) and simultaneously displayed medium rates of type II error by identifying modelled hybridization in less than half of the simulations (9/20 under *P* = 0.05; 6/20 under *P* = 0.01; see Supporting Information, Fig. S4). A concurrent result was obtained when comparing both methods and P2C2M under the empirical data of Joly & Bruneau (2006). While P2C2M identified poor model fit to the MSCM under the very gene that also indicated presence of hybridization with JML (see Supporting Information, Table S2, Fig. S5a; Joly, 2012), STEM-hy did not recover indications for hybridization in this data set (except for cases of false positives; see Supporting Information, Fig. S5b). None of the potential hybridizations inferred with JML by Joly (2012) was statistically significant at *P* = 0.05, which renders the presence of hybrids among the analysed *Rosa* species unlikely and supports the results inferred by STEM-hy. In summary, STEM-hy appears to generate more conservative inferences of hybrid detection than JML. For the present investigation, we therefore selected STEM-hy for the detection of potential hybrid taxa and, by extension, their exclusion from the data sets for AAR.

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

Potential hybrid taxa in *Tolpis*

Under the empirical data of *Tolpis*, STEM-hy inferred the participation of *Tolpis macrorhiza* (Lowe ex Hook.) DC. in hybridizations under every species tree and *Tolpis barbata* (L.) Gaertn. in more than three quarters of the species trees (Fig. 3). Other taxa are identified for hybridization in less than one fifth of the species trees, which we interpreted as false positives given the results of the corresponding simulations. JML, by comparison, indicated numerous cases of hybridization as well as a conspicuous correlation between the number of inferred hybridizations and the amount of genetic divergence per locus (see Supporting Information, Fig. S6). Under locus A19 and the ETS, hybridization was inferred among the majority of Canarian endemics as well as between the Canarian and the mainland species. The widely distributed continental and insular species *T. barbata* was inferred to partake in many of these hybridizations. Under locus B12, only a single instance of hybridization was inferred: between *T. glabrescens* and *T. succulenta*. The same hybridization event was inferred under locus A19. For the purpose of evaluating the effect on AAR, we followed the results of STEM-hy and excluded the alleles of *T. barbata* and *T. macrorhiza* from our DNA alignments and recalculated the gene and species trees.

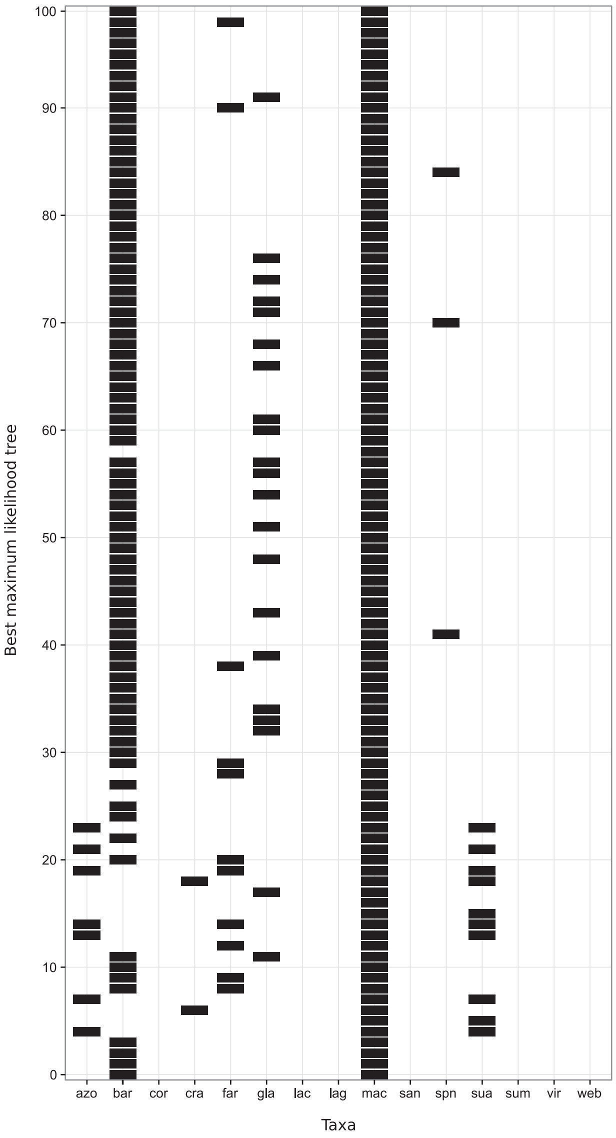

Figure 3. Results of statistical hybrid detection with STEM-hy. Each tree of the maximum likelihood species tree distribution, where a participation in hybridization was inferred, is indicated by a black bar. Taxa are abbreviated by the first three letters of their specific epithets (except for polyphyletic *T. succulenta*, which is abbreviated by 'sua' for samples from the Azores and 'sum' for samples from Madeira).

Ancestral area reconstruction

Ancestral area reconstruction by stochastic character mapping conducted prior to the exclusion of potential hybrid taxa recovered different results for different posterior tree distributions (Fig. 4; see Supporting Information, Figs S7 and S8). For clade 1, the Canary Islands were identified as the most probable ancestral distribution area under locus B12, Madeira under locus A19 as well as LS1 species trees, and the mainland under the ETS as well as LS2 species trees. For clade 2, the Azores or the Canary Islands were identified as ancestral distribution area under locus A19, the Canary Islands alone under LS1 species trees, and the mainland under locus B12, the ETS, and LS2 species trees. Ancestral area reconstruction conducted after the exclusion of *T. barbata* and *T. macrorhiza* was more homogeneous (Fig. 4). The reconstruction results indicated that each of the current habitats was a potential ancestral distribution area for clade 1, with exception of the Canary Islands under the ETS. For clade 2, the Canary Islands were inferred the most probable ancestral distribution area under each locus or locus set, except under the ETS. The reconstructions for clade 3 were consistent across different posterior tree distributions before and after the exclusion of *T. barbata* and *T. macrorhiza* in inferring the Canary Islands as the ancestral distribution area. AAR under the LS1 species trees favoured the Canary Islands as the ancestral distribution for each of these clades.

Ancestral area reconstructions under maximum likelihood and parameterized dispersal models suggested a biogeographic history of *Tolpis* that favoured dispersal from continental to island habitats. Models M2 and M3 produced the best likelihood scores in all reconstructions, except for the gene tree of B12, for which model M1 was suggested to be most likely given the sequence data (Table 3). Reconstructions under models favouring a back-dispersal to the mainland

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

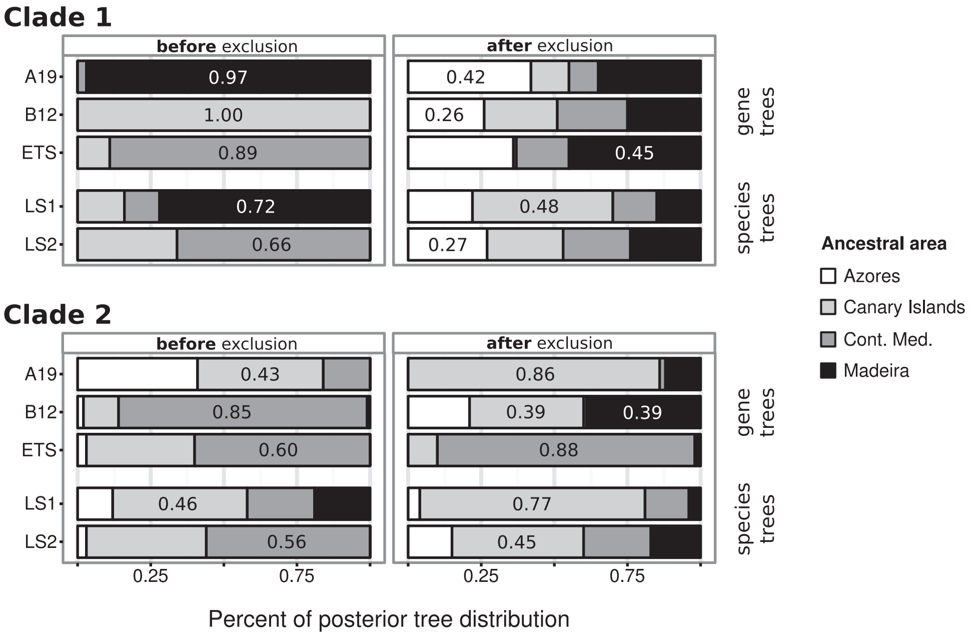

Figure 4. Results of AAR for focal clades 1 and 2. Each section of a bar chart represents the percentages of the posterior tree generations that favour a particular reconstruction. The reconstruction percentage of the most likely area per analysis is specified for each analysis. The left-hand column presents the reconstruction results prior to the exclusion of potential hybrid taxa, the right-hand column the results after their exclusion. The results for clade 3 were consistent across different loci or locus sets and thus not visualized. Abbreviations used are as in Fig. 3.

Table 3. Results of likelihood-based AAR under different models of dispersal rate to and from archipelagoes

Model		M1	M2	M3	M4	M5
Probabilities		Islands/Cont. 1.00/1.00	Islands/Cont. 1.20/0.80	Islands/Cont. 1.50/0.75	Islands/Cont. 0.80/1.20	Islands/Cont. 0.75/1.50
Gene tree – A19	Rate	10.10, 10.10	12.12, 8.08	15.15, 7.58	8.08, 12.12	7.58, 15.15
	Log *L*	–9.301	–9.058	–8.940*	–9.708	–10.187
Gene tree – B12	Rate	22.05, 22.05	26.46, 17.64	33.08, 16.54	17.64, 26.46	16.54, 33.08
	Log *L*	–9.245*	–9.250	–9.362	–9.479	–9.877
Gene tree – ETS	Rate	3.31, 3.31	3.97, 2.65	4.97, 2.48	2.65, 3.97	2.48, 4.97
	Log *L*	–8.299	–8.244*	–8.276	–8.488	–8.760
Species tree – LS2	Rate	19.79, 19.79	23.75, 15.83	29.69, 14.84	15.83, 23.75	14.84, 29.69
	Log *L*	–5.840	–5.770*	–5.816	–6.118	–6.507
Species tree – LS1	Rate	46.45, 46.45	55.74, 37.16	69.68, 34.84	37.16, 55.74	34.84, 69.68
	Log *L*	–6.117	–6.061*	–6.100	–6.375	–6.762

Asterisks indicate the reconstructions with the highest likelihood scores. Cont., Continent.

were found to fit the data worse than mainland-to-island or symmetric reconstruction models.

Ancestral area reconstructions under CTMC island models indicated support for a Canarian origin of the genus under loci A19 and B12 and, conversely, support for a continental origin of the genus under the ETS (Table 4). Likewise, the ancestral distribution area of the MRCA of all Canarian and mainland species of *Tolpis* was reconstructed to the Canary Islands under loci A19 and B12 and to the mainland under the ETS. The same results were recovered for clade 3.

DISCUSSION

Applications of statistical hybrid detection

Taxa of hybrid origin can have a considerable impact on phylogenetic tree inference, which should be

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

Table 4. Results of reconstructions under CTMC island models with an instantaneous rate matrix reflecting geographic distance between archipelagoes

Locus	Archipelago	Clade 1	Clade 2	Clade 3
Gene tree – A19	Azores	0.006 (±0.000)	0.000 (±0.000)	0.000 (±0.000)
	Continent	0.336 (±0.138)	0.000 (±0.000)	0.039 (±0.008)
	Madeira	0.124 (±0.007)	0.032 (±0.004)	0.026 (±0.004)
	Canary Islands	0.534 (±0.101)*	0.968 (±0.004)*	0.935 (±0.010)*
Gene tree – B12	Azores	0.000 (±0.000)	0.000 (±0.000)	0.000 (±0.000)
	Continent	0.002 (±0.000)	0.043 (±0.023)	0.001 (±0.000)
	Madeira	0.083 (0.065)	0.026 (±0.009)	0.000 (±0.000)
	Canary Islands	0.915 (±0.065)*	0.932 (±0.039)*	0.998 (±0.000)*
Gene tree – ETS	Azores	0.003 (±0.001)	0.000 (±0.000)	0.000 (±0.000)
	Continent	0.659 (±0.188)*	0.980 (±0.000)*	0.946 (±0.002)*
	Madeira	0.336 (±0.187)	0.007 (±0.000)	0.001 (±0.000)
	Canary Islands	0.003 (±0.000)	0.013 (±0.000)	0.053 (±0.002)

The variances of the reconstruction results are given in parentheses. Asterisks indicate the reconstructions with the highest posterior state probabilities.

considered as a source of error for cladistic analysis (McDade, 1992; Soltis *et al.*, 2008). This impact is illustrated by the multitude of phylogenetic positions that hybrid taxa can be recovered in (McDade, 1990, 1992), a phenomenon that is ultimately caused by different levels of sequence similarity and homoplasy between hybrid taxa and their parents due to the age and type of hybridization (Nieto-Feliner *et al.*, 2001; Soltis *et al.*, 2008). In particular, the inclusion of hybrid taxa formed by distantly related parents in a data set that also comprises these parents may lead to spurious phylogenetic placements, including the recovery of hybrid taxa in a clade with unrelated species (McDade, 1992). The identification of putative hybrid taxa is thus an important first step in counteracting their potential impact on phylogenetic tree inference and subsequent analyses.

In order to detect and, where necessary, exclude taxa of potential hybrid origin from our data sets prior to AAR, we evaluated and applied two statistical methods to detect reticulate patterns of allele coalescence as well as one statistical method to evaluate the fit of genetic loci to the MSCM. Each of these methods operates on DNA sequence data and evaluates a specific aspect commonly associated with hybrid speciation (Linder & Rieseberg, 2004). The methods display several advantages compared to more traditional evaluations of hybrid speciation such as artificial cross-pollination or cytogenetic experiments (Chester *et al.*, 2010). First, statistical hybrid detection is generally less expensive and less time-consuming than more traditional approaches, requiring only DNA sequence data of orthologous genomic regions. Second, statistical hybrid detection can also be applied to those hybrids that cannot be introgressed to potential parents due to hybrid sterility or the onset of parthenogenesis. Third, statistical hybrid detection is the only applicable method for investigations that rely on desiccated plant material. Thus, statistical hybrid detection is a vital tool for a wide range of investigations. However, as any statistical test, these methods contain an intrinsic statistical error that needs to be evaluated and taken into account when interpreting the results.

Evaluation of statistical hybrid detection

In this study, STEM-hy was found to be a more conservative hybrid detection method with preferable statistical power than JML (see Supporting Information, Figs S3–S5). Similar results have been reported by Heled, Bryant & Drummond (2013), who indicated that JML might suffer from elevated rates of type I error. However, our findings on the relative error rates of STEM-hy and JML are contingent on the precise dimensions of the data evaluated and do not invalidate JML as a whole. More investigation is needed to estimate the sensitivity and specificity of both methods, particularly regarding the level of genetic divergence necessary to differentiate ILS from hybridization. In particular, the results of JML point to a positive correlation between the number of reticulations detected and the amount of variation of each locus as well as the amount of genetic divergence among them. For example, the ETS has the highest number of variable sites and the largest maximum p-distance between ingroup taxa (Table 1) while displaying the largest number of inferred hybridization events under JML. The LCNM B12, by contrast, has the smallest number of variable sites and the smallest maximum *p*-distance between ingroup taxa, and it also displays the smallest number

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

of inferred hybridization events. It could, therefore, be argued that the ability of JML to differentiate ILS from hybridization may be sensitive to the overall level of genetic divergence among the input sequences. In summary, our evaluations do not provide proof for or against the integrity of either method, but allow a first estimation of their relative error rates and provide a model for future analyses.

Indications for hybridization in *Tolpis*

STEM-hy indicated that the continental and insular taxon *T. barbata* as well as the Madeiran endemic *T. macrorhiza* were participating in hybridization events (Fig. 3). The results of STEM-hy are partially consistent with those of JML, which indicated the participation of *T. barbata* in hybridizations in two loci (see Supporting Information, Fig. S6). As the only widespread and potentially invasive species of the genus, *T. barbata* displays characteristics indicative of many hybrid taxa (Rieseberg *et al.*, 2007). Similarly, *T. macrorhiza* is the only Macaronesian species of *Tolpis* that exhibits a low proportion of flowering individuals among natural populations in any given year (Crawford *et al.*, 2015); a potential connection between hybridization and floral development has been reported for Asteraceae (e.g. Bello *et al.*, 2013). The hybridization events suggested by STEM-hy are consistent with the findings of Soto-Trejo *et al.* (2013), who report that members of the Canarian clade of *Tolpis* can be readily hybridized in cultivation. For the purpose of evaluating the effect on AAR, *T. barbata* and *T. macrorhiza* were thus excluded from the empirical data sets during AAR of *Tolpis*. However, additional experimental work is warranted to confirm the true hybrid nature of both species.

The three species of *Tolpis* suggested to be hybrids by a previous molecular investigation (i.e. *T. crassiuscula*, *T. succulenta*, and *T. glabrescens*; Gruenstaeudl *et al.*, 2013) were not inferred as potential hybrids by STEM-hy. Speculation of hybrid identity in Gruenstaeudl *et al.* (2013) was largely based on the observation that species that do not inhabit the same island or even the same archipelago shared identical or highly similar ETS alleles. To conclude hybridization from this observation may have been premature, as several factors can lead to shared alleles across allopatric species. For example, nuclear ribosomal DNA is prone to concerted evolution, which streamlines allele diversity across species (Alvarez & Wendel, 2003; Poczai & Hyvönen, 2010). Likewise, ILS in recently diverged species may group together allele sequences of nonsister species during gene tree estimation (Linder & Rieseberg, 2004; Joly *et al.*, 2009). False inferences of hybridization can also occur when introgressed alleles stem from multiple, closely related sources because the independent introduction of identical alleles can obscure their inference as shared or derived (Eaton *et al.*, 2015). In summary, simply observing shared alleles across nonsister species upon gene tree estimation is insufficient to conclude that hybridization has occurred. At a minimum, a statistical evaluation to determine if the observed relationships can be explained by ILS alone is necessary (Joly, 2012).

MSCM fit and among-archipelago gene flow

When discordance between different genealogical histories exists, character reconstructions performed on species trees should allow more consistent conclusions than those on individual gene trees. However, not all genetic loci necessarily fit the MSCM, even though this is the basic assumption of most species tree estimation methods (Reid *et al.*, 2014). The identification and exclusion of loci with poor fit to the MSCM is an important prerequisite for the estimation of species trees (Kubatko, 2009) and subsequent AAR. P2C2M was developed to help identify loci with poor fit to the MSCM (Gruenstaeudl *et al.*, 2016). Prior to the empirical assessment of model fit, we conducted a set of simulations to assess the ability of P2C2M to detect cases in which poor fit is caused by hybridization. The results of these simulations indicated that P2C2M had a high statistical power when poor model fit was caused by hybridization (Fig. 2) and that it can be utilized in the identification of hybridizations under stringent significance levels.

In order to identify and, where necessary, exclude genes from the data sets prior to AAR, we utilized P2C2M on the empirical data of *Tolpis*. P2C2M indicated that both LCNM fit the MSCM (Table 2). Hence, we decided to consider only those AAR results that were generated on LS1 species trees. DNA sequences of ETS alleles, by contrast, were not found to fit the MSCM, which coincides with the observation that they display the highest relative number of variable sites and the greatest *p*-distance among the loci employed (Table 1). The high genetic variability of the ETS may be indicative of the elevated rates of nonhomologous recombination often observed in nuclear ribosomal DNA (Poczai & Hyvönen, 2010). Concerted evolution in nuclear ribosomal loci, which streamlines intraspecific allele diversity and can erase hereditary information, is often incomplete, generating novel or rearranged sequence homologs, which can facilitate non-homologous recombination (Alvarez & Wendel, 2003). It is not implausible that the lack of fit to the MSCM by the ETS is partially caused by nonhomologous recombination.

As part of this investigation, we also assessed the level of ILS across and gene flow between

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

different Macaronesian archipelagoes (see Supporting Information, Table S1). Both processes can be detrimental to the reconstruction of ancestral character states and warrant pre-analysis identification. The results indicated that the posterior distribution of LS1 species trees displayed the least amount of ILS or gene flow across archipelagoes, which corroborated our selection of LS1 species trees as the basis of AAR in *Tolpis*. Support for the scenario of reduced gene flow across archipelagoes comes from Borges-Silva *et al.* (2016), who highlight the sea as a geographical barrier to gene flow in *Tolpis*.

Ancestral area reconstruction

Our tests for hybridization, ILS, and among-archipelago gene flow highlight the need for selection among the AAR results. Thus, we only considered reconstructions that were generated (1) on data with minimal signal for hybridization, ILS, or poor model fit, and (2) after the exclusion of potential hybrid taxa. In addition, we preferred reconstructions conducted on species trees over those on gene trees in order to avoid bias through differential allele divergence. Ancestral area reconstruction on LS1 species trees inferred without *T. barbata* and *T. macrorhiza* conformed to these prerequisites. The results identified the Canary Islands as the most probable ancestral distribution area for the genus as a whole (Fig. 4), a finding supported by the observation that the majority of extant *Tolpis* species occur on this archipelago. The results under postexclusion LS1 species trees also support the Canary Islands as the most probable ancestral distribution area for the continental taxa of *Tolpis*. The reconstructions under CTMC island models corroborate both conclusions (Table 4). Given these results, it seems likely that *Tolpis* has consistently inhabited island habitats and undergone one or more dispersals from an island to a continental habitat during the course of its evolutionary history. Assuming a continental origin of the genus in the distant past, *Tolpis* underwent at least five separate dispersal events to reach its extant distribution (Fig. 5), which is the same number of dispersal steps inferred by Moore *et al.* (2002).

The application of likelihood-based reconstructions under parameterized models of dispersal inferred a biogeographic history that favoured dispersal following the currents of the north-eastern trade winds. These winds are prevalent throughout the Macaronesian region and are thought to have mediated the dispersal of many plant species (Francisco-Ortega, Jansen & Santos-Guerra, 1996). However, these results should not be taken as evidence against a back-colonization because the log likelihoods estimated under different dispersal models pertain to entire trees, not specific clades. Local deviations from the best-fitting models of dispersal are still possible. Hence, these results could be interpreted as support for a scenario in which the back-dispersal occurred between the Canary Islands and North Africa (Caujapé-Castells, 2011), which at the closest location are only 95 km apart.

Back-colonization of continent

The hypothesis of a back-colonization of the continent by *Tolpis* was first suggested by Moore *et al.* (2002), who identified a re-colonization of the mainland following the extinction of *Tolpis* in Europe and North Africa as the most parsimonious solution. No indication of a back-dispersal by *Tolpis* was found by Gruenstaeudl *et al.* (2013), but their biogeographic analyses may have been compromised by hybrid taxa. The present investigation provides support for a back-colonization of *Tolpis* by inferring that the MRCA of all extant species has likely inhabited the Canary Islands, necessitating dispersal to the continent to match the current distribution of the genus (Fig. 5). This phenomenon, termed ‘biodiversity boomerang’, has been reported from several Macaronesian plant lineages (Mort *et al.*, 2002; Allan *et al.*, 2004; Carine *et al.*, 2004; Caujapé-Castells, 2011). Biological support for such a scenario is provided by Bellemain & Ricklefs (2008), who concluded that island taxa should be capable of back-dispersal if their adaptation to island life has not resulted in reduced dispersal ability. Although dispersal in plants is predominantly unidirectional (Emerson, 2002), a back-colonization of the continent would remain consistent with this logic due the close geographic proximity of source and sink area. In fact, the species *T. barbata* inhabits the Canary Islands and most of North Africa and Mediterranean Europe, indicating that repeated dispersal may be common in this lineage.

Despite the support for a back-colonization of the continent by *Tolpis*, this scenario should be treated with caution. Like all character state reconstructions, AAR is sensitive to the precise taxon set under study. The exclusion of two species from the final data set altered the reconstruction results, although it did not alter the inference of a back-colonization (Fig. 4). When the potential hybrid taxa remain in the data set, Madeira is inferred the most probable ancestral distribution area for clade 1 and the Canary Islands for clade 2. The inference of a reverse colonization by *Tolpis* thus seems to be supported irrespective of the exclusion of potential hybrid taxa under LS1, but the inferred colonization pathway changes. Caution about the scenario of a back-colonization is also warranted given the possibility of past extinctions in *Tolpis*. Several investigations have reported likely extinctions in Macaronesian plant lineages, correlating them to climatic fluctuations in the Mediterranean during the Pliocene (e.g.

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

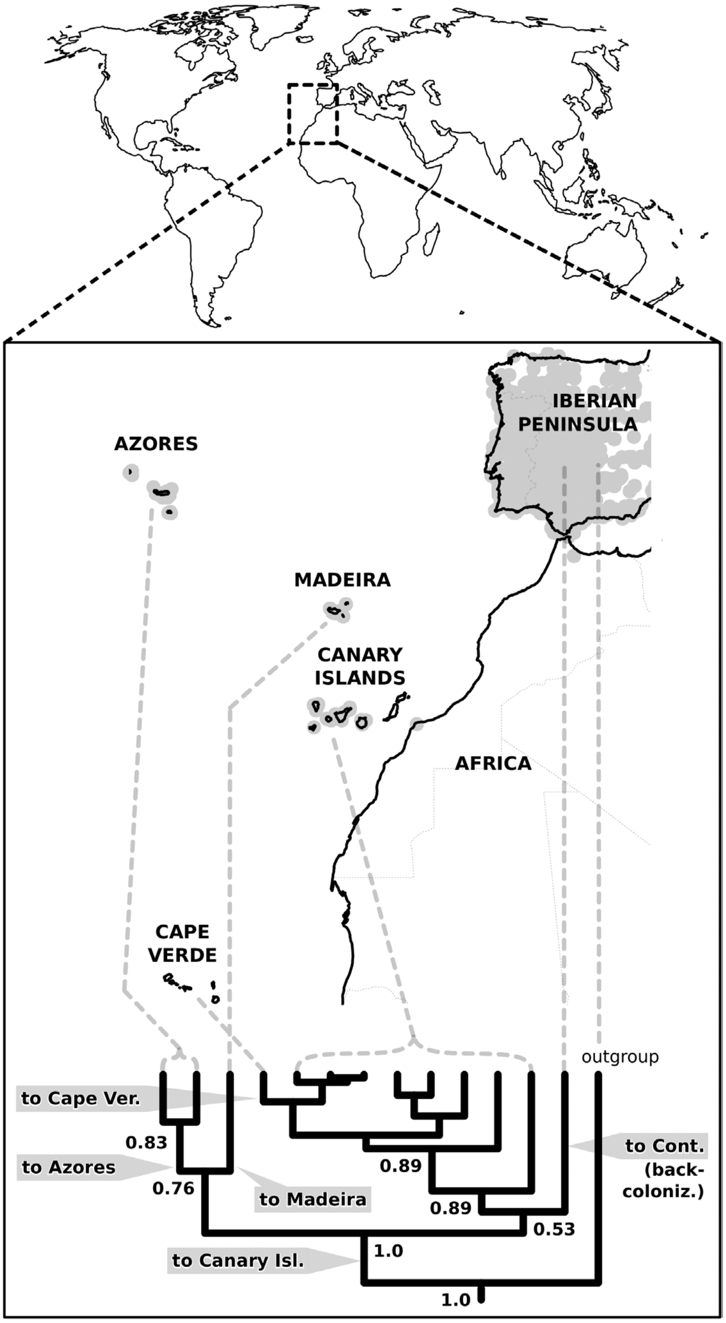

Figure 5. Phylogenetic relationships in *Tolpis* after the exclusion of potential hybrid taxa and the association with extant distribution areas. The location of the study area in a global context is displayed on the top, a detailed map that displays coastlines and country borders between the 26th meridian west, the 1st meridian east, the 43rd parallel north, and the 16th parallel north in the centre, the maximum clade credibility tree of the posterior species tree distribution inferred under locus set LS1 after the exclusion of potential hybrid taxa at the bottom of the figure. Species–area associations are visualized as dashed gray lines. Arrows next to the branches of the phylogenetic tree indicate dispersal events as inferred through AAR and the destination of each dispersal. The remaining mainland species and the outgroup were arbitrarily connected to the Iberian peninsula, given the understanding that both taxa have a much wider continental distribution. Node numbers on the phylogenetic tree indicate posterior probability values greater than 0.5. The geographic locations of entries of the keyword '*Tolpis*' in the database GBIF (http://www.gbif.org/) within the centre map are displayed as gray dots. *Tolpis* also inhabits the easternmost Canarian islands of Fuerteventura and Lanzarote, but no GBIF records for *Tolpis* on these islands seem to exist; the same is true for the Cape Verde archipelago. Cont., Continental Mediterranean; Isl., Islands.

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

Vitales *et al.*, 2014). Historical extinctions and incomplete sampling of extant taxa can be important obstacles to reconstructing colonization, as they can mislead the correct interpretation of pathways (Emerson, 2002). Although this investigation employs a comprehensive taxon set that includes all known extant species of *Tolpis*, and the diversification of Macaronesian plant lineages with similar distributions seems to have been very recent (Kim *et al.*, 2008), the existence of populations on islands where they have since become extinct cannot be discounted. Such historical populations would be particularly likely on the easternmost Canarian islands Fuerteventura and Lanzarote, which are in immediate vicinity of the North African mainland and currently inhabited by *Tolpis* only through *T. barbata*. The only moderate support for a clade of all Canarian and continental taxa of *Tolpis* (Fig. 5) could be indicative of such unsampled alleles.

Scenarios of among-archipelago gene flow

The results of the present investigation enable us to infer a possible scenario for among-archipelago gene flow in Macaronesia. Most insular plant lineages colonize specific archipelagoes only once, followed by adaptive radiations on different islands into different habitats (Emerson, 2002). Consequently, haplotypes from species endemic to specific archipelagoes are rarely shared across Macaronesian archipelagoes, but when it happens they are often interpreted as indications for among-archipelago gene flow (Pelser *et al.*, 2012; Talavera *et al.*, 2013). Reports of plant migration between Macaronesian archipelagoes (e.g. Carine *et al.*, 2004) indicate a possible scenario for such gene flow. Allopatric distributions, which constitute the main obstacle for among-archipelago gene flow, can become sympatric distributions through occasional migration, enabling introgression into the gene pool of the local endemics (Pelser *et al.*, 2012). Similarly, gene flow may be mediated by local hybridization and concluded by the local extinction of one parent. The observation that both potential hybrids identified through STEM-hy were found among the older lineages of *Tolpis* and not among the recently radiated Canarian endemics supports this scenario. Recent progress on a better resolved nuclear phylogeny of *Tolpis* (Mort *et al.*, 2015) should allow a more detailed evaluation of this scenario.

Outlook and conclusions

Next-generation sequencing and improved species tree inference methods will likely advance the statistical detection of hybrid taxa and the effort to minimize their impact on phylogeny inference and subsequent analyses. For example, the sequencing of hundreds of nuclear loci through next-generation sequencing may make it easier to exclude taxa from investigations without greatly compromising the inference of species trees (Streicher, Schulte & Wiens, 2016). Similarly, the use of phylogenomic data sets can considerably improve the inference of species trees compared to Sanger-sequenced data sets (Ruane *et al.*, 2015), which, in turn, can improve the statistical detection of potential hybrid taxa.

This investigation found that taxa of hybrid origin may have a considerable impact on ancestral area reconstruction. We advocate the application of methods that assist in the statistical detection and, where necessary, exclusion of taxa or loci with reticulate evolutionary histories prior to the estimation of species trees and the reconstruction of ancestral distribution areas. However, given that all data sets analysed here comprise only homoploid hybrid species (except for the tetraploid *T. glabrescens*), we do not recommend generalizing our conclusions to data sets of primarily allopolyploid taxa. Given significant differences in the genetic divergence of the parental species, the number and genetic diversity of alleles in allopolyploids may be substantially different from those of homoploid hybrids (Paun *et al.*, 2009). Additional research is needed to corroborate the hybrid nature of the taxa indicated here and to evaluate the statistical power of the applied methods on allopolyploid taxa. Based on reconstructions adjusted for putative hybrid taxa, we conclude that the plant genus *Tolpis* has likely had a time-consistent distribution in island habitats and that it likely originated on the Canary Islands. Our results also support the scenario that *Tolpis* is part of a distinct group of Macaronesian plant lineages that have experienced a back-dispersal to the continent.

ACKNOWLEDGMENTS

This investigation was supported by the Blake Centennial Professorship in Systematic Botany of The University of Texas at Austin to R.K.J., by a start-up package from The Ohio State University to B.C.C., graduate student research awards from the American Society of Plant Taxonomists and the Botanical Society of America to M.G. and the Plant Resources Center at The University of Texas at Austin. We also thank three anonymous reviewers for their helpful comments. This work was supported in part by an allocation of computer resources from the HPC Service of ZEDAT of the Freie Universität Berlin and from the Ohio Supercomputer Center.

References

Allan GJ, Francisco-Ortega J, Santos-Guerra A, Boerner E, Zimmer EA. 2004. Molecular phylogenetic evidence for

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

the geographic origin and classification of Canary Island *Lotus* (Fabaceae: Loteae). *Molecular Phylogenetics and Evolution* **32:** 123–138.

Alvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29:** 417–434.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22:** 341–345.

Bellemain E, Ricklefs RE. 2008. Are islands the end of the colonization road? *Trends in Ecology & Evolution* **23:** 461–468.

Bello MA, Álvarez I, Torices R, Fuertes-Aguilar J. 2013. Floral development and evolution of capitulum structure in *Anacyclus* (Anthemideae, Asteraceae). *Annals of Botany* **112:** 1597–1612.

Borges-Silva L, Sardos J, Menezes de Sequeira M, Silva L, Crawford D, Moura M. 2016. Understanding intra and inter-archipelago population genetic patterns within a recently evolved insular endemic lineage. *Plant Systematics and Evolution* **302:** 367–384.

Carine MA, Russell SJ, Santos-Guerra A, Francisco-Ortega J. 2004. Relationships of the Macaronesian and Mediterranean floras: molecular evidence for multiple colonizations into Macaronesia and back-colonization of the continent in *Convolvulus* (Convolvulaceae). *American Journal of Botany* **91:** 1070–1085.

Caujapé-Castells J. 2011. Jesters, red queens, boomerangs and surfers: a molecular outlook on the diversity of the Canarian endemic flora. In Bramwell D, Caujapé-Castells J, eds. *The biology of island floras.* Cambridge: Cambridge University Press, 284–324.

Chester M, Leitch AR, Soltis PS, Soltis DE. 2010. Review of the application of modern cytogenetic methods (FISH/GISH) to the study of reticulation (polyploidy/hybridisation). *Genes* **1:** 166–192.

Clarkson JJ, Knapp S, Garcia VF, Olmstead RG, Leitch AR, Chase MW. 2004. Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. *Molecular Phylogenetics and Evolution* **33:** 75–90.

Crawford DJ, Anderson GJ, Borges-Silva L, Menezes de Sequeira M, Moura M, Santos-Guerra A, Kelly JK, Mort ME. 2015. Breeding systems in *Tolpis* (Asteraceae) in the Macaronesian islands: the Azores, Madeira and the Canaries. *Plant Systematics and Evolution* **301:** 1981–1993.

Crawford DJ, Mort ME, Archibald JK. 2013. *Tolpis santosii* (Asteraceae: Cichorieae), a new species from La Palma, Canary Islands. *Vieraea* **41:** 163–169.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29:** 1969–1973.

Eaton DAR, Hipp AL, González-Rodríguez A, Cavender-Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution* **69:** 2587–2601.

Emerson BC. 2002. Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. *Molecular Ecology* **11:** 951–966.

Fougère-Danezan M, Joly S, Bruneau A, Gao XF, Zhang LB. 2015. Phylogeny and biogeography of wild roses with specific attention to polyploids. *Annals of Botany* **115:** 275–291.

Francisco-Ortega J, Jansen RK, Santos-Guerra A. 1996. Chloroplast DNA evidence of colonization, adaptive radiation, and hybridization in the evolution of the Macaronesian flora. *Proceedings of the National Academy of Sciences of the United States of America* **93:** 4085–4090.

Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* **49:** 652–670.

Gompert Z, Buerkle CA. 2016. What, if anything, are hybrids: enduring truths and challenges associated with population structure and gene flow. *Evolutionary Applications* **9:** 909–923.

Gruenstaeudl M. 2016. WARACS: wrappers to automate the reconstruction of ancestral character states. *Applications in Plant Sciences* **4:** 1500120.

Gruenstaeudl M, Reid NM, Wheeler GL, Carstens BC. 2016. Posterior predictive checks of coalescent models: P2C2M, an R package. *Molecular Ecology Resources* **16:** 193–205.

Gruenstaeudl M, Santos-Guerra A, Jansen RK. 2013. Phylogenetic analyses of *Tolpis* Adans. (Asteraceae) reveal patterns of adaptive radiation, multiple colonization and interspecific hybridization. *Cladistics* **29:** 416–434.

Heled J, Bryant D, Drummond AJ. 2013. Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evolutionary Biology* **13:** 44.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27:** 570–580.

Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. *Systematic Biology* **52:** 131–158.

Huelsenbeck JP, Rannala P, Masly JP. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288:** 2349.

Jarvis CE. 1980. *Systematic studies in the genus Tolpis Adanson*. Unpublished D. Phil. Thesis, University of Reading.

Joly S. 2012. JML: testing hybridization from species trees. *Molecular Ecology Resources* **12:** 179–184.

Joly S, Bruneau A. 2006. Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from *Rosa* in North America. *Systematic Biology* **55:** 623–636.

Joly S, Heenan PB, Lockhart PJ. 2014. Species radiation by niche shifts in New Zealand's rockcresses (*Pachycladon*, Brassicaceae). *Systematic Biology* **63:** 192–202.

Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist* **174:** E54–E70.

Jones KE, Reyes-Betancort JA, Hiscock SJ, Carine MA. 2014. Allopatric diversification, multiple habitat shifts, and hybridization in the evolution of *Pericallis* (Asteraceae), a Macaronesian endemic genus. *American Journal of Botany* **101:** 637–651.

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

Kim SC, McGowen MR, Lubinsky P, Barber JC, Mort ME, Santos-Guerra A. 2008. Timing and tempo of early and successive adaptive radiations in Macaronesia. *PLoS One* **3:** e2139.

Kubatko LS. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology* **58:** 478–488.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* **25:** 971–973.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50:** 913–925.

Linder CR, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany* **91:** 1700–1708.

Maddison WP, Maddison DR. 2011. *Mesquite: a modular system for evolutionary analysis, Version 2.75.* Available at: http://mesquiteproject.org

McDade L. 1990. Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. *Evolution* **44:** 1685–1700.

McDade L. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* **46:** 1329–1346.

Moody ML, Rieseberg LH. 2012. Sorting through the chaff, nDNA gene trees for phylogenetic inference and hybrid identification of annual sunflowers (*Helianthus* sect. *Helianthus). Molecular Phylogenetics and Evolution* **64:** 145–155.

Moore MJ, Francisco-Ortega J, Santos-Guerra A, Jansen RK. 2002. Chloroplast DNA evidence for the roles of island colonization and extinction in *Tolpis* (Asteraceae: Lactuceae). *American Journal of Botany* **89:** 518–526.

Mort ME, Crawford DJ, Kelly JK, Santos-Guerra A, Menezes de Sequeira M, Moura M, Caujapé-Castells J. 2015. Multiplexed-shotgun-genotyping data resolve phylogeny within a very recently derived insular lineage. *American Journal of Botany* **102:** 634–641.

Mort ME, Soltis DE, Soltis PS, Francisco-Ortega J, Santos-Guerra A. 2002. Phylogenetics and evolution of the Macaronesian clade of Crassulaceae inferred from nuclear and chloroplast sequence data. *Systematic Botany* **27:** 271–288.

Nieto-Feliner G, Fuertes-Aguilar J, Rosselló J. 2001. Can extensive reticulation and concerted evolution result in a cladistically structured molecular data set? *Cladistics* **17:** 301–312.

Nieto-Feliner G, Rosselló JA. 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution* **44:** 911–919.

Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* **48:** 612–622.

Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* **53:** 673–684.

Paun O, Forest F, Fay MF, Chase MW. 2009. Hybrid speciation in angiosperms: parental divergence drives ploidy. *The New Phytologist* **182:** 507–518.

Pelser PB, Abbott RJ, Comes HP, Milton JJ, Moeller M, Looseley ME, Cron GV, Barcelona JF, Kennedy AH, Watson LE, Barone R, Hernandez F, Kadereit JW. 2012. The genetic ghost of an invasion past: colonization and extinction revealed by historical hybridization in *Senecio*. *Molecular Ecology* **21:** 369–387.

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Systematic Biology* **50:** 580–601.

Poczai P, Hyvönen J. 2010. Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. *Molecular Biology Reports* **37:** 1897–1912.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computational Applications to the Biosciences* **13:** 235–238.

Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. *Tracer, Version 1.6.* Available at: http://beast.bio.ed.ac.uk/Tracer

Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. *Systematic Biology* **63:** 322–333.

Rieseberg LH, Kim SC, Randell RA, Whitney KD, Gross BL, Lexer C, Clay K. 2007. Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica* **129:** 149–165.

Ronquist F. 2004. Bayesian inference of character evolution. *Trends in Ecology & Evolution* **19:** 475–481.

Ronquist F, Sanmartin I. 2011. Phylogenetic methods in biogeography. *Annual Review of Ecology Evolution and Systematics* **42:** 441–464.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61:** 539–542.

Ruane S, Raxworthy CJ, Lemmon AR, Lemmon EM, Burbrink FT. 2015. Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. *BMC Evolutionary Biology* **15:** 1–14.

Soltis DE, Mavrodiev EV, Doyle JJ, Rauscher J, Soltis PS. 2008. ITS and ETS sequence data and phylogeny reconstruction in allopolyploids and hybrids. *Systematic Botany* **33:** 7–20.

Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. *Annual Review of Plant Biology* **60:** 561–588.

Soto-Trejo F, Kelly JK, Archibald JK, Mort ME, Santos-Guerra A, Crawford DJ. 2013. The genetics of self-compatibility and associated floral characters in *Tolpis* (Asteraceae) in the Canary Islands. *International Journal of Plant Sciences* **174:** 171–178.

Streicher JW, Schulte JA 2nd, Wiens JJ. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in Iguanian lizards. *Systematic Biology* **65:** 128–145.

Talavera M, Navarro-Sampedro L, Ortiz PL, Arista M. 2013. Phylogeography and seed dispersal in islands: the case of *Rumex bucephalophorus* subsp. *canariensis* (Polygonaceae). *Annals of Botany* **111:** 249–260.

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017

Timme RE, Simpson BB, Linder CR. 2007. High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S-26S ribosomal DNA external transcribed spacer. *American Journal of Botany* **94:** 1837–1852.

Tippery NP, Les DH. 2011. Phylogenetic relationships and morphological evolution in *Nymphoides* (Menyanthaceae). *Systematic Botany* **36:** 1101–1113.

Toepel M, Lundberg M, Eriksson T, Eriksen B. 2011. Molecular data and ploidal levels indicate several putative allopolyploidization events in the genus *Potentilla* (Rosaceae). *PLoS Currents* **1:** RRN1237.

Vitales D, Garnatje T, Pellicer J, Vallès J, Santos-Guerra A, Sanmartín I. 2014. The explosive radiation of *Cheirolophus* (Asteraceae, Cardueae) in Macaronesia. *BMC Evolutionary Biology* **14:** 1–15.

Xiang QP, Wei R, Shao YZ, Yang ZY, Wang XQ, Zhang XC. 2015. Phylogenetic relationships, possible ancient hybridization, and biogeographic history of *Abies* (Pinaceae) based on data from nuclear, plastid, and mitochondrial genomes. *Molecular Phylogenetics and Evolution* **82 Pt A:** 1–14.

Yu Y, Degnan JH, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics* **8:** e1002660.

Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences of the United States of America* **111:** 16448–16453.

Zhang JQ, Meng SY, Allen GA, Wen J, Rao GY. 2014. Rapid radiation and dispersal out of the Qinghai-Tibetan Plateau of an alpine plant lineage *Rhodiola* (Crassulaceae). *Molecular Phylogenetics and Evolution* **77:** 147–158.

Zhu S, Degnan JH, Goldstien SJ, Eldon B. 2015. Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees. *BMC Bioinformatics* **16:** 292.

Zwickl DJ. 2006. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Unpublished D. Phil. Thesis, University of Texas at Austin.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article:

Figure S1. Example of the impact of hybrid or allopolyploid taxa on the reconstruction of ancestral character states.
Figure S2. Two examples of gene and species trees inferred from simulated data as employed to evaluate the statistical power of STEM-hy and JML in detecting potential cases of hybridization.
Figure S3. Results of simulation analyses to evaluate the statistical power of STEM-Hy to detect cases of hybridization among a set of taxa.
Figure S4. Results of simulation analyses to evaluate the statistical power of JML to detect cases of hybridization among taxa of specific loci.
Figure S5. Results of the evaluation and comparison of the statistical power of JML and STEM-hy when applied to the empirical data set of Joly & Bruneau (2006).
Figure S6. Results of statistical hybrid identification with JML on the empirical data sets of *Tolpis*.
Figure S7. Results of AAR prior to the exclusion of *Tolpis* species that were indicated to be potential hybrids by STEM-hy.
Figure S8. Results of AAR after the exclusion of *Tolpis* species that were indicated to be potential hybrids by STEM-hy.
Table S1. Results of an evaluation of the level of incomplete lineage sorting and among-archipelago gene flow in *Tolpis* via Bayesian and likelihood-based hypothesis testing.
Table S2. Results of analyses with P2C2M on the empirical data set of Joly & Bruneau (2006).

Downloaded from https://academic.oup.com/biolinnean/article-abstract/121/1/133/2915831/Statistical-hybrid-detection-and-the-inference-of
by Ohio State University-Moritz Law Library user
on 12 September 2017